Developing and Measuring IS Scales Using Item Response Theory

Completed Research Paper

Thomas Rusch
Assistant Professor
Competence Center for Empirical Research Methods
WU (Vienna University of Economics and Business Administration)
Welthandelsplatz 1, D4
1020 Vienna
Austria
Thomas.Rusch@wu.ac.at

Patrick Mair
Lecturer
Department of Psychology
Harvard University
33 Kirkland Street
Cambridge, MA 02138
United States of America
mair@fas.harvard.edu

Paul Benjamin Lowry
Associate Professor
Department of Information Systems
City University of Hong Kong
P7912, Academic 1
88 Tat Chee Avenue, Kowloon
Hong Kong, People's Republic of China
Paul.Lowry.PhD@gmail.com

Horst Treiblmaier
Full Research Professor
Logistikum Steyr
University of Applied Sciences Upper Austria
Wehrgrabengasse 1-3
4400 Steyr
Austria
Horst.Treiblmaier@fh-steyr.at

Abstract

Information Systems (IS) research frequently uses survey data to measure the interplay between technological systems and human beings. Researchers have developed sophisticated procedures to build and validate multi-item scales that measure latent constructs. Most studies use classical test theory (CTT), which suffers from several theoretical shortcomings. We discuss these problems and present item response theory (IRT) as a viable alternative. Subsequently, we use the CTT approach as well as Rasch models (a class of restrictive IRT models) to develop a scale for measuring the hedonic aspects of websites. The results illustrate how IRT can not only be successfully applied in IS research but also provide improved results over CTT approaches.

Keywords: Item Response Theory, Rasch Model, Partial Credit Model, Measurement, Hedonism, Hedonic Information Systems, Classical Test Theory

Introduction

Social science research and Information Systems (IS) research have produced a wealth of empirical papers that use survey data either to create new measurement scales or to apply previously validated scales to measure constructs. Additionally, new research is being published in IS journals on how to improve current methods. In most cases, the authors rely on fundamental principles that have been developed and refined in classical test theory (CTT) over decades. The applicability of CTT to IS has rarely been questioned (for an exception see Rossiter 2011).

Psychometricians such as Thurstone (1925), Rasch (1960), or Birnbaum (1968) have formulated different statistical models to achieve the measurement of latent traits. Two main approaches for measuring continuous latent traits have emerged: CTT (Gulliksen 1950) and item response theory (IRT) (Birnbaum 1968; Rasch 1960), with the former gaining widespread popularity. Even though the first IRT models were introduced decades ago, their application in scholarly research is still limited (Borsboom 2006). Today, most research papers utilizing IRT can be found in psychology and educational testing, and, at the same time, these papers are slowly gaining popularity in marketing research. In recent years, several publications have clearly shown the advantages of this measurement approach and thus have sparked new interest (Salzberger and Sinkovics 2004). IS researchers, however, have virtually ignored IRT.

Typically, when researchers measure latent variable(s), they strive to find a "good" subset of items that allows for reliable, highly informative, and objective measurement of the underlying construct. However, such measurement often demands fundamental statistical requirements that cannot be sufficiently met by CTT approaches. IRT, on the other hand, offers tools that can meet these requirements. Specifically, in this paper, we use and demonstrate Rasch models. Although Rasch models are a restrictive form of IRT models in terms of item selection and model fit, they provide a number of properties that are advantageous for scale development and substantive research based on these scales. Other IRT models are more flexible and therefore better suited for modeling but lack these properties.

To demonstrate the applicability of Rasch models to IS research, we have developed a scale of hedonism based on hedonic IS. IS research shows the increasing importance of hedonism in studying IS use (Deng et al. 2010; Lin and Bhattacherjee 2010; Lowry et al. 2013; van der Heijden 2004; Wakefield and Whitten 2006). Instead of seeing individuals as rational beings that actively process great amounts of information before making shopping decisions, researchers such as Hirschman and Holbrook (1982) highlight the hedonic, aesthetic, and symbolic nature of the consumption process. With the advent of technical systems that offer advanced multimedia capabilities, and particularly the World Wide Web, utilitarian and hedonic concepts have begun to intermingle. For this reason, hedonic motivation was recently added as a key component to the unified theory of acceptance and use of technology (Venkatesh et al. 2012) and is the foundation of the proposed hedonic-motivation system adoption model (HMSAM) (Lowry et al. 2013).

For our purpose, which is to illustrate the applicability of IRT by developing a scale to measure hedonism complying with the Rasch criteria, we initially created an item base (i.e., website attributes) that was as broad as possible. The goals of the comparative analyses with CTT and IRT were to find items that measure hedonism as a latent construct unidimensionally, to provide information on how these items actually measure the underlying construct, and to see how much information we can obtain from them.

Concepts and Assumptions of CTT and IRT

The common approach in scale construction is known as CTT (Lord and Novick 1968). Its basic statistical model is $X=T+E$, where X denotes the observed overall score, T the true overall score (which can be seen as a latent construct), and E the measurement error. Hence, the observed score is assumed to be a linear function of the underlying true score. This very restrictive assumption cannot be tested (Fischer 1974). Obviously, if one wants to model every single answer, this identity can be assumed to hold on the item-specific level. The idea of this linear relationship can be generalized to include the idea of factor models for one or more latent variables with different loadings (or regression slopes) per item, such as in a common factor model or confirmatory factor model (e.g., Bartholomew and Knott 1999). The key issue in CTT and its generalizations is that the observed scores are linearly regressed on the latent constructs.

A key critique of CTT is that the right-hand side is unknown; thus, to meet the equation, T and E can be

chosen arbitrarily. Thus, this equation is a tautology rather than a statistical model (Fischer 1974). Regardless, researchers typically use reliability coefficients based on this basic expression. Reliability is defined as $\rho^2(X, T) = \sigma^2(T)/\sigma^2(X)$. Since T is unknown, we cannot compute its variance $\sigma^2(T)$. Hence, additional assumptions are needed in terms of measurement equivalence of test splitting; hence, reliability is commonly estimated as internal consistency by means of Cronbach's α (1951). However, α indirectly includes the correlations between the items and therefore is inherently a measure of a linear relationship between items. In the case of α and other correlation-based reliability measures of CTT, the scores must be on an interval scale; otherwise, any correlation-based reliability will not be invariant.

The assumptions of linearity and having a metric scale are central to CTT. For example, when one constructs a questionnaire with these assumptions, the whole process of item selection is based on correlation coefficients. The square root of the reliability is expressed as $r(X, T)$ and the discriminatory power of item i (i.e., whether item i measures something nearly identical such as the test composite score) as $r(X_i, X)$. Items that are highly correlated are retained and items that are weakly correlated with other items are eliminated. Oftentimes, Cronbach's α, which is a property of a set of items, is even used for item selection through the deletion of items with a low score. If factor models are employed, their estimates are also based on correlations.

A major problem with the assumption of a linear relationship is that, if the latent trait is assumed to be on an interval scale, researchers treat the observed scores or sum scores as if they were metric as well. Such is usually not the case if the questionnaire uses indicators such as Likert-type scales, nominal indicators such as qualitative statements, and (if we are strict) bounded metric scales. Here, we summarize three main shortcomings of CTT, as addressed mainly in psychology (e.g., Borsboom 2006; Fischer 1974; Hambleton and Jones 1993; Weiss and Davison 1981) and marketing contexts (Salzberger 2007).

First, the assumption of a linear relationship between the latent and observed scores is very restrictive and is known not to necessarily represent the empirical reality when it comes to psychological constructs (see, Fischer and Formann 1982). Additionally, assuming such a linear function with different locations of the items implies that, for certain values of the latent trait, no probability to answer in a certain way is defined. This implication is undesirable, because it restricts the span of the latent variable. If such a linear relationship is assumed, it is not congruent with the idea of different item locations. The same problem arises if different item discrimination (i.e., different slopes of the linear function) is assumed. If one postulates this linear relationship, all items must have the same discrimination and location, which is called the assumption of τ-equivalent measures (Lord and Novick 1968).

Second, in CTT, the true score or factor score that is actually of interest cannot be estimated directly, but only via additional assumptions regarding the item-specific true scores. A scoring rule is implicitly assumed to be correct, but its adequateness cannot be tested. Furthermore, the sums of observed scores are often taken as an estimate of the person's or the item's location. This calculation equates the expected true values with the sum of the observed scores. However, it is possible that a person with a lower score in reality will have a higher position on the latent trait. Such could be the case, for example, if a person with a lower position picks an answer at random. Therefore, using the sum of observed scores is not necessarily appropriate for measuring this empirical reality, since doing so runs counter to the idea of measurement as a mapping of empirical relations to numerical relations. It is possible that a scoring rule other than the sum of observed scores is the better choice, but when we use CTT, we cannot find out if this is the case.

Third, parameters such as reliability, discrimination, location, or factor structures depend on the sample being used, which implies different reliabilities as well as different factor loadings of an item set for homogeneous and heterogeneous samples, respectively (Fischer 1974). Whenever these values are calculated, they apply only to the sample at hand and are unbiased for the population of interest only if the sample is a true random sample and representative for the population of interest (e.g., Embretson and Reise 2000; Fischer 1974). They depend on the distribution of the latent traits for a given population and vice versa. If we want to estimate the location of a person on a latent trait, that value depends on the sample of items used for measurement and on the other persons who are assessed. Depending on the reference population, a person will also have a different position on the latent trait even if the random sample is representative. Thus, such measurement can never be objective.

These shortcomings are inherent to CTT and cannot be resolved within the theory. However, we are quick to point out that CTT is not an incorrect approach per se. Instead, it is the method of choice when working

with assessment on metric scales. The same applies to generalizations such as confirmatory factor analysis or structural equation models. CTT provides a rich framework to conduct analyses if two key assumptions hold: (1) if it is theoretically justifiable that the observed scores lie on a metric scale (even for Likert-type items) and (2) if the observable and latent variables are linearly related.

Item Response Theory (IRT)

A possible remedy for the problems of CTT with non-metric non-linear relationships is to adopt a different toolbox, often collected under the umbrella term "item response models." While we approach the presentation of IRT from the point of view of practical application and statistical models, we should point out here that the implications of using IRT for measurement go far beyond that, especially concerning the underlying measurement paradigm. For these important aspects, we refer the interested reader to Salzberger and Koller (2013) and Borsboom (2005) as a starting point.

To illustrate the underlying rationale of IRT models, we will assume the simplest case of dichotomous items, coded with one and zero. Let β_i denote the location (a parameter connected with an item) of an item i (i=1,...,K), i.e., the higher its value, the less the probability of scoring 1. More accurately, β_i is the value on the latent trait where scoring 1 has a probability of 0.5 for this item. Let θ_v denote the position of person v (v=1,..,N) on the latent trait. If $\beta_i = \theta_v$, the probability of scoring 1 is 0.5 for that person. We therefore get a (0,1) persons × items data matrix $\mathbf{X}$ of dimension $N \times K$. The item response patterns X_i and person-response patterns X_v are indicators for β_i and θ_v. Other than in confirmatory factor analysis (CFA), neither causal nor distributional assumptions of latent traits need to be imposed (but can be). The patterns X_i and X_v are still on an ordinal level, but β_i and θ_v lie on an interval scale.

Let B denote a matrix S x K of S different item parameters in columns (e.g., location) and the respective values of these parameters for each of the K items in rows. Let Θ denote the matrix that gives the positions on the latent dimensions that are underlying the behavior in a certain situation. Each column refers to a specific latent dimension and its entries to the location of people on that latent dimension if presented with all the K items. The basic functional relation is then $P(\mathbf{X} = X) = f(\mathrm{B}, \Theta)$. Different IRT approaches exist in terms of the number of item-related parameters or person-related parameters. For instance, in addition to the location parameters β_i, the researcher might wish to allow for item-discrimination parameters α_i (how well the item discriminates between the positions of two people on the latent trait) or parameters that capture picking answers at random γ_i, which in some situations might be more realistic. For both item- and person-related parameters, it is the case that if a multidimensional construct is used, every person could have a different latent trait value on all dimensions, and every item might measure each trait to a certain degree. The traits can also be correlated.

Depending on the degree of parameterization and the overall goal of the analysis, two conceptual approaches exist in IRT: (1) **Item selection approach**: The aim is to find items for which a certain (restrictive) IRT model holds. These items will then exhibit various properties of these models, such as the "fairness" of comparing people, sample independence of estimates, different discrimination ability, or heterogeneous locations. The most important of these models, the Rasch models, are parsimonious in terms of the item-related parameters. These models allow for objective measurement (Rasch 1960).

(2) **Modeling approach**: If researchers are not primarily interested in selecting items in a very restrictive manner to form a scale, but instead wish to analyze a person's response behavior and therefore the scale and its items, then they would take into account higher parameterized models (e.g., 2PL or 3PL models), multidimensional models, semi- and nonparametric models, and/or models with covariates (de Boeck and Wilson 2004).

Advantages of IRT

As was noted earlier, when the goal is to measure a latent construct, CTT and related methods can lead to serious problems. Borsboom (2006) writes that "in an alternative world, where CTT was never invented, the first thing a researcher, who has proposed a measure for a theoretical attribute, would do is to spell out the nature and the form of the relationship between the attribute and its putative measures" (p. 429). That is exactly what IRT does, and in doing so overcomes several limitations of CTT.

First, the linear relation between indicators and a categorical response, which is assumed in CTT, is usually

not appropriate. In IRT, nonlinear relationships are used. Such a nonlinear function is more general and subsumes a linear relationship. The nonlinear function that relates the probability of observing a certain response to an individual item with the latent trait is called the item response function (IRF) or item characteristic curve (ICC). This function enables flexible specifications of the theoretical relationship between the underlying trait and the items, given response format (e.g., dichotomous or polytomous), contexts, or theoretical assumptions about the response process (e.g., dimensionality). Additionally, from an empirical point of view, the higher flexibility of IRT models allows for a close fit to be achieved between a function and the data. For example, if the real relationship is linear, it is possible to fit a near linear function with the two-parameter logistic model (Birnbaum 1968), whereas the opposite is not true.

The most popular function employed is the logistic function, e.g., in the one-parameter logistic model (Rasch 1960) or two- and three-parameter logistic models (Birnbaum 1968). The former assumes constant discrimination of items and unidimensional measurement; the latter additionally allows for different discrimination or even a "pick-answer-at-random" parameter. Multidimensional generalizations exist for most of these models (Adams et al. 2007). Although the model employed is just as arbitrary as assuming a linear function, IRT models allow assessment of the adequateness by means of statistical goodness-of-fit tests and fit indices.

Second, IRT allows the analysis to be carried out on a response pattern level where the researcher can use the full amount of available information, rather than on an aggregated correlation level. IRT models usually allow the researcher to find an estimate of the underlying latent trait value that incorporates all the available information, as well as an appropriate scoring rule to adequately represent the empirical relationships given a certain probability model (called the "sufficient statistic"). In particular cases, they even allow the use of the sum scores as a sufficient statistic (i.e., in Rasch measurement). Other IRT models have different sufficient statistics (e.g., the sum of scores weighted with the discrimination parameter for the two-parameter logistic model). In CTT, the sufficient statistic for the true score is the sum of scores only if the scores are normally distributed. This cannot be the case if dichotomous or polytomous items are used.

Third, certain IRT models (i.e., Rasch models) allow the estimation of item parameters independently of the sample that has been used. This possibility means that, if the model holds for the population, any conclusions derived from the estimated parameters are valid for the population regardless of what sample has been used. For example, if a Rasch model holds in a population consisting half of females and half of males, using only males to estimate parameters is perfectly valid and leads on average to the same estimate as if only females were used. A bigger sample size improves estimation accuracy, but assessing the two most able people will not bias the results. Thus, a representative sample is not necessary if the model holds in the population. Any comparison of people within this sample will be sample-independent.

Fourth, when selecting items to construct a scale, IRT enables the researcher to select items that are in accordance with a desired model, most commonly Rasch models. Out of a pool of possible items for the scale, the ones that conform to a Rasch model might be selected to ensure that its measurement properties apply. This selection is guided by the usage of statistical tests as well as graphical procedures. In CTT, items are often chosen following some rules of thumb on underlying approximate and sample dependent measures. It is possible (and common) that both approaches lead to similar or even equivalent scales, but this does not have to be the case. Even if they do, the researcher can be considered fortunate that the statistically inappropriate model yielded the same scale as a more correct approach did.

Fifth, concepts such as reliability, internal consistency, and validity are applicable to IRT models. Reliability and internal consistency of a set of items will be high if a unidimensional IRT model holds, because all items measure the same trait. Then, all items are homogeneous in terms of the trait(s) they measure. IRT models also frequently allow the researcher to gain more information regarding how well an individual item measures. Specifically, IRT allows the researcher to assess the precision of measurement or standard error of every single item. Therefore, confidence intervals for an estimate of the latent trait value can be calculated, and these intervals will depend on the accuracy of the estimation of the latent trait values. This is in accordance with the empirical findings that measurement in middle regions of the latent trait is more accurate than in the extreme regions (Kubinger 2003). Here, the difference from CTT is that the accuracy is not assumed to be constant. CTT assumes a common standard error for all items, calculated as an average precision across the whole sample based on a reliability measure. Therefore, every item is assumed to measure every latent trait value with the same accuracy. Once again, this assumption is very restrictive and included in IRT approaches.

Sixth, with IRT, it is possible to obtain detailed information at an item and person level simultaneously. Each item i and each person v is assigned one or more parameters (i.e., location of item β_i and position on trait θ_v) that allows for a probabilistic analysis of the response behavior. Item and person parameters lie on an interval scale, which makes it possible to interpret the distances between items and persons on Θ. This is especially notable when the observed responses are on a non-metric scale. Certain IRT models actually enhance the scale level. If the items and persons are on the same scale, then statements about the response probability of person v on item i can be achieved, i.e., it is possible to predict the behavior of a person or object to a certain item.

Additional Properties of Rasch Models

Rasch (1960) reasoned about requirements to be fulfilled so that a specific proposition can be regarded as scientific. His conclusion was that a basic requirement is the objectivity of comparisons (Rasch 1961), and he formulated the epistemological theory of specific objectivity (SO): *objective* because any comparison of a pair of parameters (items/persons) should be independent of any other parameters or comparisons; *specifically objective* because the comparison made is relative to some specified frame of reference (Andrich 1988). That is, under SO, two persons v and w with latent trait positions θ_v and θ_w are comparable independently from the remaining persons in the sample and independently from the item subset with which they were presented. In turn, two items i and j with locations β_i and β_j are comparable independently from the remaining items in the subset and independently from the people in the sample (Mair and Hatzinger 2007a). To achieve this, very strict requirements are applied to these IRT models, which lead to scales that exhibit extraordinary measurement qualities.

Rasch (1960) presented a probabilistic model that can be used to study the response behavior of individuals on dichotomous items. It poses a logistic relation between the position θ_v of a person v and the probability for a "1" response on item i. Each item gets a location parameter β_i. The formal representation, which is known as Rasch model is

$$P(X_{vi}=1|\theta_v,\beta_i)=p_{vi}=\frac{\exp(\theta_v-\beta_i)}{1+\exp(\theta_v-\beta_i)},$$

where $P(X_{vi}=1)$ is the probability that person v answers in category "1" on item i. (We will drop the indices in the following sections.)

Figure 1 represents the basic ideas in terms of the ICC for two items. The probability to answer "1" is depicted on the ordinate: the abscissa displays the latent trait value. The probability for an observed score of 1 (solid line) and 0 (dotted line) for item i as a function of the latent trait, the ICC, is shown. For item j, only the ICC for score 1 is depicted. The vertical solid lines represent the item location on the common scale of item and latent trait (i.e., P(X=1)=0.5). For item j, a higher latent trait value is needed to achieve the same probability to observe score 1 than for item i. For item i with β=-0.3, both the probability to observe "0" (dotted line) and the probability to observe "1" (solid line) as a logistic function of the underlying latent trait value are shown. These two lines intersect at P(X=1)=0.5, which is by definition the "location" β of item i. This value can also be interpreted as the threshold at which it becomes more likely to score 1 than 0. One can see that the higher the position of a person on the latent trait, the higher the probability to score 1 becomes, and vice versa. For item j, only the probability to score 1 is shown because P(X=0)=1-P(X=1). This item has a higher location (β_j=1) than item i, which means the probability to score 1 is lower than for item i for any given latent trait value. It is noteworthy in this case that both items have the same discrimination, i.e., "slope" of the logistic curve. Therefore, Rasch models do not allow the logistic curves to cross. In a two-parameter logistic model (2PL-model), this would not be the case.

Because of the functional relationship, no matter what the latent trait value is, a score is always defined. Additionally, we can assess the whole latent trait as long as we have enough items that are different in terms of their locations. Another interesting issue is that, in the middle region, around the item's location, measurement is actually linear, but for extreme regions, the item is not able to distinguish well between people. It can also be seen that peoples' abilities and item difficulties lie on the same scale. Furthermore, the intersection point cuts the latent trait into a region that corresponds to score 0 and score 1, respectively, which means that the model also maps the dichotomous responses onto a metric scale.

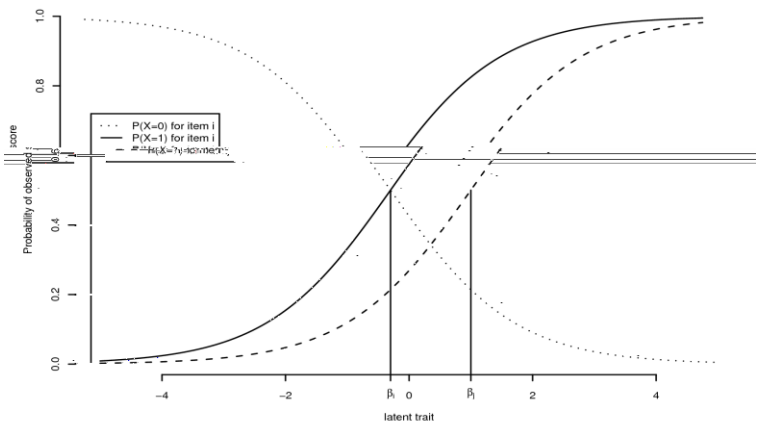

Figure 1: Representation of a Dichotomous Rasch Model for Two Items

Figure 1 also shows how restrictive the dichotomous Rasch model actually is. It assumes that (a) there is only one underlying latent trait, (b) a constant discrimination parameter for different items exists, (c) that the ICC is a logistic function, and (d) that the probability for a certain score depends solely on item location and person location on the latent trait. The main prerequisite, however, is that "local independence" is fulfilled. This prerequisite is shared by many other IRT models. It means that they assume independence of the item responses conditional on the person's latent trait value. The interdependency and statistical relationship between observations on different items is solely due to the position of the person on the latent trait. Therefore, concepts that are based on correlations between items, as in CTT, are seen as spurious. For example, a highly aggressive person will often agree to items that ask for violent behavior, whereas a pacifist will not. If the latent trait value is held constant, any relationship between items disappears. As opposed to CTT, all of these restrictions can be tested. There are model tests for every assumption, the most important and popular one being the likelihood-ratio (LR) test proposed by Andersen (1973), which is based on (person) sample splits.

Because of its restrictiveness, the Rasch model is not flexible enough for modeling purposes, but if all items conform to it, this model has some remarkable features, as described above. To achieve model fit, one would eliminate items that contradict at least one of the Rasch model assumptions. Item selection can be conducted by different means, for example, by residual-based item fit statistics (Smith 2004) or Wald tests (Glas and Verhelst 1995). Those items that remain in the final homogeneous item subset measure the latent construct in an objective manner as proposed by Rasch (1960).

In many practical situations, dichotomous item responses are too restrictive, especially in social science research, where Likert-scales are commonly used for assessing individuals' attributes. For such polytomous items, the model outlined above can be generalized. One popular extension is the partial credit model (PCM) (Masters 1982). Embretson and Reise (2000) give an extensive discussion of parameter interpretations and relations between polytomous IRT models. Since we focus on the item selection approach, we limit further explanations to the PCM.

Using the PCM, all of the properties and assumptions of the dichotomous Rasch model still apply. Every ordinal item i with m_i as the number of categories is described by h-1 cumulative intersection parameters β_{ih}that map the categories onto the latent trait. The PCM can thus be regarded as an adjacent-categories logit model (Tuerlinckx and Wang 2004). For each category, there is a probability to score in this category as a function of the latent trait; it estimates log-odds for a certain category h with respect to category h-1. An important issue therefore is the interpretation of the item-category parameters β_{ih}. These parameters are often transformed into category intersection parameters δ_{ij} with $j = 0 \ldots m_i$. If we estimate the PCM, the item categories are converted into intersection parameters as $\delta_{i0} = -\beta_{i0};\ \delta_{i1} = \beta_{i0} - \beta_{i1};\ \delta_{i2} = \beta_{i1} - \beta_{i2}$. The parameters δ_{ij} refer to the points on the latent trait where the ICCs intersect. Based on these

intersection parameters, we can compute item location parameters v_i in terms of $v_i = m_i^{-1} \sum_{j=0}^{m_i} \delta_{ij}$. Within the context of item selection to construct a scale, the main focus is on the item (-category) parameters that we can estimate independently from person parameters if Rasch models are applied. In this case, we are not primarily interested in the estimation of θ. Our aim is to establish a homogeneous subset of items that allows for a specific objective measurement of a latent construct. To score persons, we would need a useful scale having a wide range of items in terms of their location.

Measuring Hedonic Information Systems

Here, we demonstrate the applicability of the Rasch Model in IS by constructing a scale that allows us to measure hedonic IS. This scale is supposed to measure only one latent dimension. Although this dimension can also be a higher-order factor, all items of the scale should be able to assess it. To provide a more useful demonstration, we have created two scales: one using a CTT approach, and one using the PCM. The latter analysis will serve as a guideline to illustrate how scales can be constructed using an IRT approach.

We used several steps to collect and clean the data. To ensure content validity, we used a panel of seven experts to generate a list of properties that are important for websites. The design of our study was straightforward: we simply asked the experts to "name attributes that can be used to describe websites." In order to facilitate this process, we showed them various randomly selected sites from different categories, and they simply described their properties. We designed this phase as a brainstorming session, with the major objective being to come up with as many attributes as possible without any evaluation or rating. Subsequently, we used the same panel of experts to group the items they chose and to filter out synonyms, which resulted in a total of 26 items.

After performing ten preliminary tests to ensure the items were comprehensible, we conducted an online survey in which a convenience sample of 291 Internet users rated the importance of those attributes for online customer portals. We did not give any example of websites in order to avoid a bias that may have been caused by different levels of familiarity with a site. We used a 5-point scale with a range from zero ("not important") to four ("very important") to assess the significance of the single attributes. Therefore, the data matrix **X**, which we use for all subsequent analyses, consists of 291 subjects and 26 items. Since we conducted all surveys in German, we used a translation and back-translation approach to ensure semantic consistency. When no agreement could be reached, we consulted an additional translator.

CTT Analysis

We used R (R Development Core Team 2007) with the packages "psych" (Revelle 2009) for exploratory factor and reliability analysis, and "sem" (Fox 2009b) for confirmatory factor analysis. To calculate the polychoric correlations, we used the package "polycor" (Fox 2009a). In CTT, the first problem is to find out how many latent factors may be underlying our items. For our purpose, we aimed for one factor only. No clear-cut solution exists—there are various criteria to help choose the number of factors. We selected principal axis factoring of the polychoric correlation matrix and used the scree plot criterion as well as the Very Simple Structure (VSS) criterion (Revelle and Rocklin 1979) to determine the number of factors as well as Revelle's (1979) β and McDonald's (1999) ω_h to check the general factor solution.

There was an enormous drop in explained variance after the first factor is extracted. Therefore, the scree plot as well as the VSS criterion (with a maximum value of 0.86 at factor 1 for complexity 1) suggest extracting only one factor. To verify this, we conducted a CFA with one common factor. All items were allowed to load freely on the common factor. We found that the loading of one item "plain" was not significantly different from 0 (α=.05); thus, we deleted the item and refitted the CFA. The fit indices suggest that the model does not fit the data very well (RMSEA=0.111, CFI=0.725, ω_h=0.72, β=0.67). Since our aim was to derive a single scale within a CTT framework, we followed scale construction conventions but were quick to point out that we did not endorse these conventions or the used measures, as there are limitations associated with them (Diamantopoulos et al. 2012; Drolet and Morrison 2001). Accordingly, we proceeded with deleting items that had small loadings on the factor from the one-factor solution, until we found an acceptable fit or until the fit became worse again. The best fit was achieved after the deletion of the items "multimedia based," "beautiful," "modern," "interactive," "customized," "personal," "tasteful," and "plain" (RMSEA=0.106, CFI=0.835, ω_h =0.8, β=0.68), which left 18 items. Even this "best" CTT model does not fit

the data well. Nevertheless, the selected items are suited for a unidimensional measurement of hedonic information systems within a CTT framework. Note that our scale has a Cronbach's α of 0.91, and all loadings are significant, a fact that obscures the bad fit of the unidimensional solution. Table 1 shows the one-factor solution before and after item selection.

Table 1: Results of the confirmatory factor analysis with polychoric correlation

	Factor 1 before selection		Factor 1 after selection	
Item	Loading	SE	Loading	SE
surprising	0.696	0.052	0.715	0.052
intriguing	0.563	0.055	0.557	0.056
inspiring	0.589	0.055	0.563	0.055
playful	0.694	0.052	0.702	0.052
animated	0.659	0.053	0.650	0.054
multimedia based	0.506	0.056	-	-
funny	0.687	0.053	0.697	0.052
entertaining	0.762	0.051	0.767	0.051
provocative	0.537	0.056	0.564	0.055
motivating	0.591	0.055	0.561	0.056
beautiful	0.504	0.056	-	-
exciting	0.723	0.052	0.717	0.052
coltish	0.767	0.050	0.785	0.050
modern	0.423	0.058	-	-
emotional	0.703	0.052	0.697	0.052
colorful	0.619	0.054	0.618	0.054
full of action	0.716	0.052	0.735	0.051
humorous	0.742	0.051	0.748	0.051
challenging	0.667	0.053	0.670	0.053
interactive	0.335	0.059	-	-
customized	0.399	0.058	-	-
personalized	0.424	0.058	-	-
tasteful	0.370	0.059	-	-
suitable for children	0.478	0.057	0.475	0.057
creative	0.580	0.055	0.544	0.056
Chi-Square (df)	1264.6 (275)		578.71 (135)	
CFI	0.725		0.835	
RMSEA	0.111		0.106	
NFI	0.676		0.797	
NNFI	0.700		0.814	
SRMR	0.008		0.06	
Cronbach's α	0.91		0.91	

IRT Analysis

We performed all computations with the eRm package (Mair and Hatzinger 2007a; b) in R, which uses CML estimation and allows for computation of the test statistics previously described. To achieve a final set of items, we used the following steps (cf. Figure 2): (1) estimate the item and subject parameters of the PCM; (2) compute item-fit statistics based on the residuals; (3) eliminate the item with the smallest *p*-value; and (4) compute the *LR*-test for different person sub splits (if *LR* is significant, go back to step (1) and proceed with item elimination). Otherwise, the procedure stops, and we obtain the final model.

When the data did not fit the PCM[1], we eliminated items successively and re-fitted the model. The result

[1] It is a peculiarity of the item selection approach that data are actually fitted to a model, not the other way round (as is mainly the case in statistics). Such is the case because Rasch models are restrictive IRT models

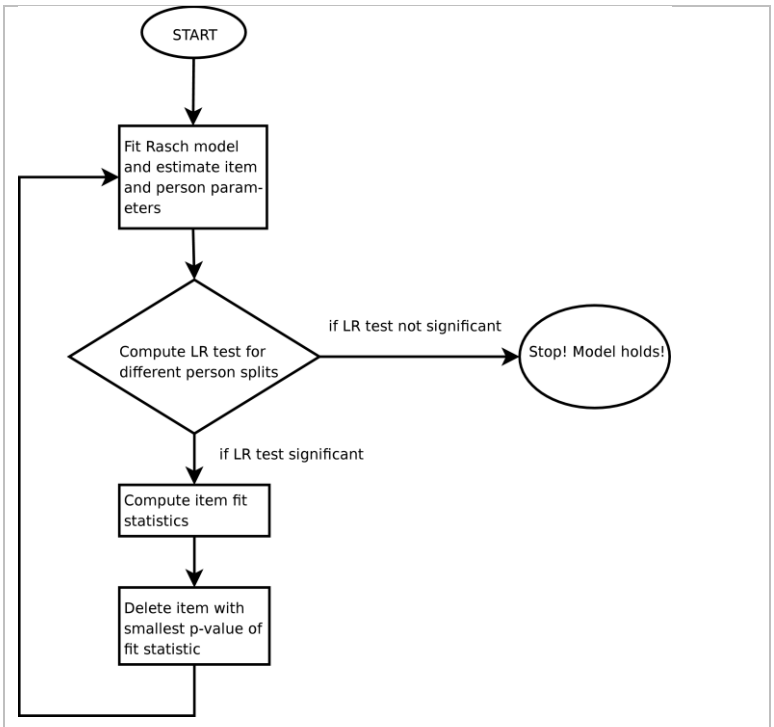

Figure 2: Item Selection Approach with Rasch Models

was a set of homogeneous items that comply with the restrictive Rasch criteria. This means that the items all measure the same latent trait (i.e., unidimensional), that the sum of scores is the appropriate measure of the underlying latent trait, and that the estimated parameters are sample-independent (i.e., specific objectivity holds) if the model holds in the population. The reason for fitting the *LR*-test after each step is that this statistic, which is a global model test, evaluates model fit of the whole item set. Item-fit statistics are residual based and compare a theoretical probability with an observed integer value. Thus, this criterion is only suitable for indicating which items should be eliminated, but not for assessing model fit.

We started our analysis with the same total set of 26 items that we used in the previous section. We eliminated, based on the procedure described above, the following items in this order: "plain," "suitable for children," "interactive," "customized," "personalized," "multi-media based," "modern," "tasteful," "beautiful," "creative," "provocative," "inspiring," "intriguing," "colorful," and "animated." The remaining 11 items were appropriate for scaling the hedonic aspects of websites within a Rasch framework. Ranked from the largest to the smallest *p*-value of the item fit statistics, they are "frisky," "humorous," "entertaining," "full of action," "exciting," "surprising," "emotional," "playful," "challenging," "funny," and "motivating." For this set of items, we applied a small simulation of 40 *LR*-tests by means of person-splits (2-group random-splits and 3-group random splits). The corresponding *p*-values were not significant at the 5% level in 95% of the cases (38 of 40). Therefore, the PCM can be assumed to hold for this scale.

Table 2 shows the item location parameters v_i and the category intersection parameters δ_{ij} for the final item subset. These parameter sets allow for a detailed interpretation of each item. The final items are heterogeneous in their locations on the "importance for measuring hedonism" latent scale—ranging from "motivating" (v_i=-0.582) on the left-hand side of the continuum up to "frisky" (v_i=1.155) on the right-hand side. Note that when measuring hedonism it is irrelevant which items from this pool are chosen because all

and would not hold without taking special care when we develop the items. Other IRT models allow for the conventional statistical approach of model fitting.

Table 2: Location and threshold parameters of items selected in the Rasch model

	Location	Threshold1	Threshold2	Threshold3	Threshold4
surprising	0.55017	−0.3765	0.8126	0.55318	1.21139
playful	0.31256	−0.6884	0.39266	0.20719	1.3388
funny	0.31414	−0.5191	0.23827	0.13385	1.40349
entertaining	−0.3055	−1.1942	0.0799	−0.6724	0.56484
motivating	−0.5825	−1.3400	−0.4936	−1.0895	0.5932
exciting	−0.0518	−1.1906	0.08804	−0.2669	1.16215
frisky	1.15526	0.46667	1.40648	0.66995	2.07795
emotional	0.53599	−0.5029	0.41212	0.84531	1.38936
full of action	0.64045	−0.3493	0.48214	0.8303	1.5987
humorous	0.08177	−0.6286	0.3733	−0.4814	1.06374
challenging	0.25191	−0.5727	0.39598	−0.0284	1.21273

of them comply with the Rasch model and thus are appropriate for measuring hedonism. Also, since raw scores are the appropriate scores for a Rasch-type model, all items/people with the same raw score would get the same parameter and thus lie on the same position on Θ, the latent trait. The Rasch model holds, which assures that it is eligible (in terms of its sufficient statistic, so information is not lost) and objective in summing up the scores of the final item subset, either for people or for items.

Consequently, in this case, the estimated location parameters can be seen as ranked in the following way: the higher the location, the more important the item is considered to be for hedonism. Location parameters allow for the interpretation of differences in importance according to the construct hedonism. For instance, the difference in item location between "emotional" and "funny" (0.535-0.314=0.22) is approximately 2.4 times as much as between "full of action" and "surprising" (0.64-0.55=0.09). Thus, the latter are 2.4 times more similar in terms of the amount of the construct the items represent than the former.

The category intersection parameters δ_{ij} (also called threshold parameters) denote the points on the latent continuum Θ at which the category characteristic curves (CCC) intersect. Figure 3 shows several examples of the underlying CCCs. Each line visualizes the probability of observing a certain response as a function of the latent trait values. For the item "emotional," the categories 0 and 1 intersect at a value of -0.502. This implies that as long as we have an estimated hedonism score below -0.502, the probability of a zero score on this item will be higher than for any other category. As long as a person thinks that the importance of this item for measuring hedonism is between [-0.502;0.414], the person will most probably select a response of "1," but need not do so. A person may choose "4," but such a response is much less likely than "1." Thus, unlike in CTT, the researcher can interpret the results probabilistically.

The items "emotional" and "full of action" possess a "regular" behavior in terms of increasing intersection parameters (as the category increases), i.e., $\delta_{i0} < \delta_{i1} < \delta_{i2} < \delta_{i3}$. This monotonicity property is not given for all the other items, as is shown at the bottom of Figure 3. It is especially striking that $\delta_{i2} < \delta_{i1}$ for the item "playful". That does not imply that there are not enough subjects with a score of three; rather, it shows that, conditional on the hedonism score, the probability for a response in category 2 is lower throughout, compared to responses on other categories. This behavior can frequently be observed for neutral or middle categories. The PCM assigns intervals on the unidimensional latent trait to a certain score. These intervals must be ordered as well. If such is not the case, then, in contrast to the assumption taken during item development, category three cannot be mapped in this way. No interval is assigned mainly to this category, which suggests that something about this category is not in accordance with the ordinal ordering of categories. For scale development, this indicates that the categories of items that display this behavior are not ordinal, but that there are 4 categories that are ordered, and that the middle category is different. One could now either change the Likert-type scale to a 4-point scale (with no middle category, since it measures

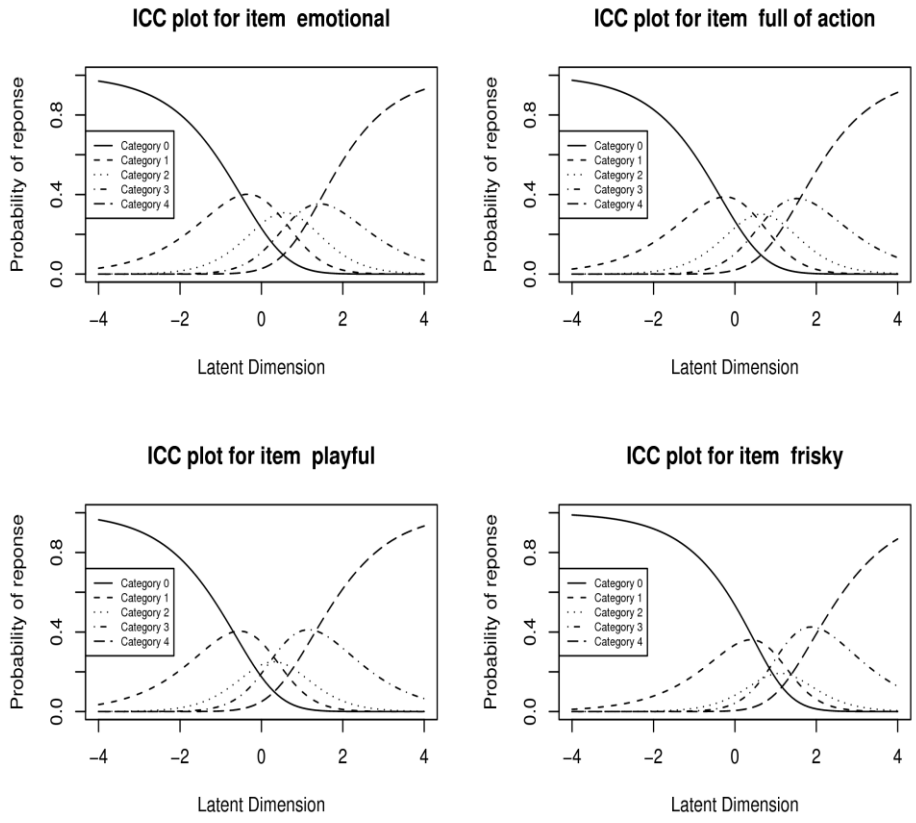

Figure 3: ICCs for the Five Categories

something else, probably "undecided"), use an IRT model that assumes the scores to be only nominal scaled, or a model that estimates the four "regular" categories as ordinal and the middle category as "nominal" (Bartholomew and Knott 1999). Such insight is not possible with CTT.

To assess how well an item measures its latent trait values, IRT employs the concept of information of items, which tells the researcher how much information an individual item can give about certain latent trait values. For what follows, we use the formula by Samejima (1970)to calculate information. In doing so, we can see which region of the latent trait is measured well by which items or which items are redundant (e.g., two items measure nearly the same latent trait region, but one has higher information). We can also add up these item information values to a joint information value if we want to compare different scales to measure hedonism. Figure 4 shows a plot of the item information (top) and scale information (bottom) as a function of the underlying latent trait values. It can be seen that the item "exciting" measures the latent trait in an interval similar to the rest of the items, but has less information. Thus, if one wants to reduce the number of items further, this would be a potential candidate for removal. The items "frisky" and "motivating" are very important, for they have high information for rather low or high positions on the latent trait, respectively. From the scale information plot, it can be seen that this scale has slightly more information for values of θ higher than 0.

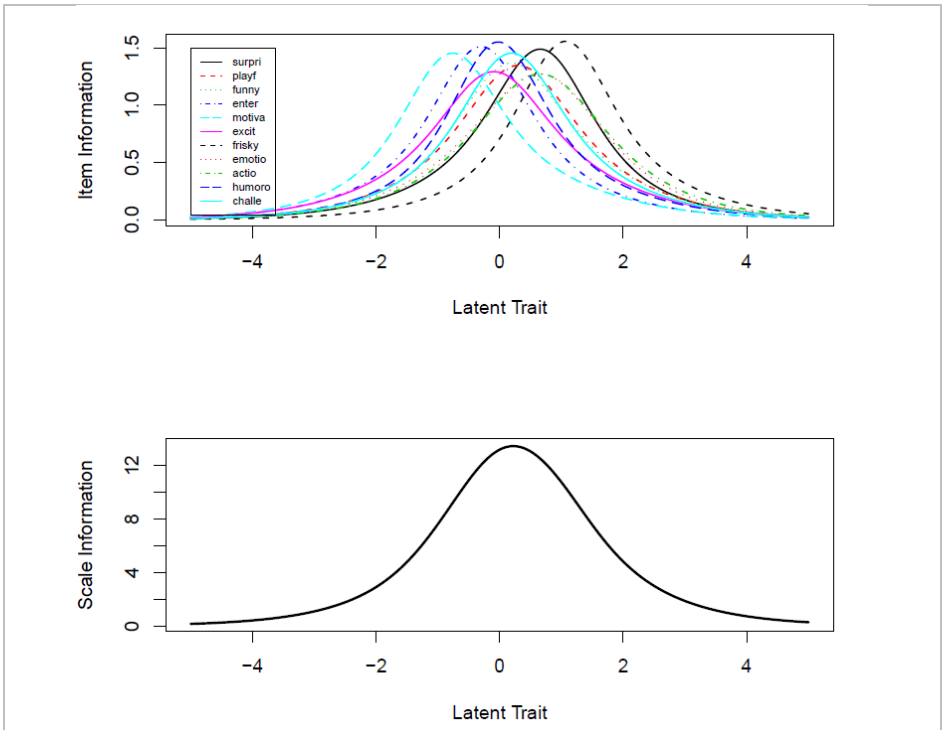

Figure 4: Information Curves for All Items (Top) and the Aggregated Information of the Scale (Bottom) as a Function of the Latent Trait Values

Discussion and Conclusion

We introduced the IRT paradigm of measurement in an IS context of hedonic websites and illustrated the practical applicability of a probabilistic framework to measure latent constructs. We did so by means of polytomous Rasch models and found attributes suitable for characterizing hedonic aspects of websites. We derived and compared scales both under the IRT and the CTT paradigms and concluded that the scale derived under IRT not only has the same reliability and fewer items then the CTT scale, but also provides additional insight. Namely, IRT provides more information about the individual scale and its items and embeds the scale construction process and the derived scholarly results into a strong theoretical and epistemological context of measurement. The IRT analysis not only allows for probabilistic statements about an individual's answering behavior but also indicates (a) how well the expression of the latent construct subjects can be assessed, (b) how well the overall latent construct can be assessed, and (c) how the individual items scale the individuals. Table 3 compares the applicability of CTT and IRT.

Even though IRT models were developed decades ago and are well-founded statistically, they are not widely applied in IS research, in which IS scholars mainly rely on the CTT paradigm. Consequently, the manifest potential of IRT has been largely overlooked by most IS researchers. By correctly applying this method, scholars can gain new insights about the content domain of frequently used constructs.

Another promising approach for future IRT applications lies in the development and implementation of multi-dimensional IRT models, which map items and persons simultaneously onto multiple correlated dimensions (von Davier and Carstensen 2007). As soon as the conceptual understanding disseminates outside the psychometricians' community and social science researchers learn how to apply this method and interpret the results, these researchers will possess a powerful measurement instrument that overcomes several shortcomings of CTT. Further research is needed in order to assess existing IS scales using CTT and IRT, e.g., by measuring their criterion or predictive validity. Additionally, we suggest that frequently used IS scales be re-evaluated with those different measurement paradigms.

Table 3: Comparison of the CTT and IRT properties (extended from Hambleton and Jones 1993)		
	CTT	IRT
Model Type	Linear	Nonlinear
Scale Level of Item	Metric	Categorical
Level of Application	Item set	Individual item
Assumptions	Weak (easier to meet with data)	Strong (more difficult to meet with data)
Item-Ability Relationship	Not specified (usually linear function)	Item characteristic function
Ability Indicator (Range)	Test scores/estimated true score (restricted to range of raw scores)	Person parameter ($-\infty, +\infty$)
Invariance of Item & Person Statistics	No	Yes (if model holds)
Reliability/Internal Consistency	Estimated reliability	Model inherent
Assumptions Testable?	No	Yes

References

Adams, R.J., Wu, M.L., and Carstensen, C.H. 2007. "Application of Multivariate Rasch Models in International Large-Scale Educational Assessments," in *Multivariate and Mixture Distribution Rasch Models: Extensions and Applications,* M. Davier and C.H. Carstensen (eds.), New York, NY: Springer, pp. 271-280.

Andersen, E.B. 1973. "A Goodness of Fit Test for the Rasch Model," *Psychometrika* (38:1973), pp. 123-140.

Andrich, D. 1988. *Rasch Models for Measurement*, Newbury Park, CA: Sage.

Bartholomew, D.J., and Knott, M. 1999. *Latent Variable Models and Factor Analysis*, London, UK: Hodder Arnold.

Birnbaum, A. 1968. "Some Latent Trait Models," in *Statistical Theories of Mental Test Scores,* F.M. Lord and M.R. Novick (eds.), Reading, MA: Addison-Wesley, pp. 395-479.

Borsboom, D. 2005. *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*, Cambridge, MA: Cambridge University Press.

Borsboom, D. 2006. "The Attack of the Psychometricians," *Psychometrika* (71:3), pp. 425-440.

Cronbach, L.J. 1951. "Coefficient Alpha and the Internal Structure of Tests," *Psychometrika* (16:1951), pp. 297-334.

de Boeck, P., and Wilson, M. 2004. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*, New York, NY: Springer.

Deng, L., Turner, D.E., Gehling, R., and Prince, B. 2010. "User Experience, Satisfaction, and Continual Usage Intention of IT," *European Journal of Information Systems* (19:1), pp. 60-75.

Diamantopoulos, A., Sarstedt, M., Fuchs, C., Wilczynski, P., and Kaiser, S. 2012. "Guidelines for Choosing between Multi-Item and Single-Item Scales for Construct Measurement: A Predictive Validity Perspective," *Journal of the Academy of Marketing Science* (40:3), pp. 434-449.

Drolet, A.L., and Morrison, D.G. 2001. "Do We Really Need Multiple-Item Measures in Service Research?," *Journal of Service Research* (3:3), pp. 196-204.

Embretson, S.E., and Reise, S. 2000. *Item Response Theory for Psychologists*, Mahwah, NJ: Lawrence Erlbaum.

Fischer, G.H. 1974. *Einführung in Die Theorie Psychologischer Tests [Introduction to Mental Test Theory]*, Bern, Germany: Huber.

Fischer, G.H., and Formann, A.K. 1982. "Some Applications of Logistic Latent Trait Models with Linear Constraints on the Parameters," *Applied Psychological Measurement* (6:1982), pp. 397-416.

Fox, J. 2009a. "Polycor: Polychoric and Polyserial Correlations R Package Version 0.7-7," retrieved: November 6, 2012, from http://cran.r-project.org/web/packages/polycor/.

Fox, J. 2009b. "Sem: Structural Equation Models. R Package Version 0.9-16," retrieved: April 18, 2012, from http://CRAN.R-project.org/package=sem.

Glas, C., and Verhelst, N. 1995. "Tests of Fit for Polytomous Rasch Models," in *Rasch Models. Their Foundation, Recent Developments and Applications,* G.H. Fischer and I.W. Molenaar (eds.), New York, NY: Springer, pp. 325-352.
Gulliksen, H. 1950. *Theory of Mental Tests*, New York, NY: Wiley.
Hambleton, R.K., and Jones, R.W. 1993. "Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development," *Educational Measurement* (12:3), pp. 38-47.
Hirschman, E.C., and Holbrook, M.B. 1982. "Hedonic Consumption: Emerging Concepts, Methods and Propositions," *Journal of Marketing* (46:3), pp. 92-101.
Kubinger, K.D. 2003. "Adaptives Testen [Adaptive Testing]," in *Schlüsselbegriffe Der Psychologischen Diagnostik [Key Terms of Psychological Assessment],* K.D. Kubinger and R.S. Jäger (eds.), Basel, Switzerland: Beltz.
Lin, C.-P., and Bhattacherjee, A. 2010. "Extending Technology Usage Models to Interactive Hedonic Technologies: A Theoretical Model and Empirical Test," *Information Systems Journal* (20:2), pp. 163-181.
Lord, F.M., and Novick, M.R. 1968. *Statistical Theories of Mental Test Scores*, Reading, MA: Addison Wesley.
Lowry, P.B., Gaskin, J., Twyman, N., Hammer, B., and Roberts, T.L. 2013. "Proposing the Hedonic-Motivation System Adoption Model (Hmsam) to Increase Understanding of Adoption of Hedonically Motivated Systems," *Journal of the Association for Information Systems* (forthcoming).
Mair, P., and Hatzinger, R. 2007a. "Cml Based Estimation of Extended Rasch Models with the Erm Package in R," *Psychology Science* (49:2007), pp. 26-43.
Mair, P., and Hatzinger, R. 2007b. "Extended Rasch Modeling: The Erm Package for the Application of Irt Models in R," *Journal of Statistical Software* (20:9), pp. 1-20.
Masters, G.N. 1982. "A Rasch Model for Partial Credit Scoring," *Psychometrika* (47:1982), pp. 149-174.
McDonald, R.P. 1999. *Test Theory: A Unified Treatment*, Mahwah, NJ: L. Erlbaum Associates.
R Development Core Team 2007. "R: A Language and Environment for Statistical Computing," retrieved: April 18, 2012, from http://www.r-project.org/.
Rasch, G. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*, Copenhagen, Denmark: Danish Institute for Educational Research.
Rasch, G. 1961. "On General Laws and the Meaning of Measurement in Psychology," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA: University of California Press, pp. 321-333.
Revelle, W. 1979. "Hierarchical Clustering and the Internal Structure of Tests," *Multivariate Behavioral Research* (14:1), pp. 59-69.
Revelle, W. 2009. "Psych: Procedures for Psychological, Psychometric, and Personality Research. R Package Version 1.0-67," retrieved: April 18, 2012, from http://CRAN.R-project.org/package=psych.
Revelle, W., and Rocklin, T. 1979. "Very Simple Structure: An Alternative Procedure for Estimating the Optimal Number of Interpretable Factors," *Multivariate Behavioral Research* (14:1979), pp. 403-414.
Rossiter, J.R. 2011. "Measurement for the Social Sciences. The C-Oar-Se Method and Why It Must Replace Psychometrics," *European Journal of Marketing* (45:11/12), pp. 1561-1588.
Salzberger, T. 2007. "Scientific Measurement of Latent Variables in Marketing Research: An Alternative Framework," Unpublished, Vienna University of Economics and Business Administration, Vienna, Switzerland.
Salzberger, T., and Koller, M. 2013. "Towards a New Paradigm of Measurement in Marketing," *Journal of Business Research* (66:9), pp. 1307-1317.
Salzberger, T., and Sinkovics, R.R. 2004. "Reconsidering the Problem of Data Equivalence in International Marketing Research," *International Marketing Review* (23:4), pp. 390-417.
Samejima, F. 1970. "Estimation of Latent Ability Using a Response Pattern of Graded Scores," *Psychometrika* (35:1), pp. 139-139.
Smith, R.M. 2004. "Fit Analysis in Latent Trait Measurement Models," in *Introduction to Rasch Measurement,,* E.S. Smith and R.M. Smith (eds.), Maple Grove, MN: JAM Press, pp. 73-92.
Thurstone, L.L. 1925. "A Method of Scaling Psychological and Educational Tests," *Journal of Educational Psychology* (16:1925), pp. 433-451.
Tuerlinckx, F., and Wang, W. 2004. "Models for Polytomous Data," in *Explanatory Item Response Models:*

A Generalized Linear and Nonlinear Approach, P. de Boeck and M. Wilson (eds.), New York, NY: Springer, pp. 75-110.
van der Heijden, H. 2004. "User Acceptance of Hedonic Information Systems," *MIS Quarterly* (28:4), pp. 695-704.
Venkatesh, V., Thong, J.Y.L., and Xu, X. 2012. "Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology," *MIS Quarterly* (36:1), pp. 157-178.
von Davier, M., and Carstensen, C.H. 2007. *Multivariate and Mixture Distribution Rasch Models: Extensions and Applications*, New York, NY: Springer.
Wakefield, R.L., and Whitten, D. 2006. "Mobile Computing: A User Study on Hedonic/Utilitarian Mobile Device Usage," *European Journal of Information Systems* (15:3), pp. 292-300.
Weiss, D.J., and Davison, M.L. 1981. "Test Theory and Methods," *Annual Review of Psychology* (32:1981), pp. 629-658.